**Dokumentasi Big Data**

**Deskripsi Dataset :**

Dataset (tugas4.txt) berisi berita yang diambil dari website kompas.com dan detik.com. Dataset memiliki 30 record yang mana masing-masing record memiliki 3 atribut. Atribut pertama merupakan id berita (dimulai dari 1 hingga 30). Atribut kedua merupakan tanggal berita tersebut diterbitkan. Atribut ketiga merupakan text berita yang akan diproses selanjutnya.

ID berita 1-10 dikumpulkan pada hari Minggu, 15 Mei 2016 di detik.com,  
ID berita 11-20 dikumpulkan pada hari Senin, 16 Mei 2016 di detik.com,  
dan ID berita 21-30 dikumpulkan pada hari Jumat, 13 Mei 2016 di kompas.com.

**Proses DataMining :**

Pertama, seluruh teks pada isi berita diganti menjadi huruf kecil seluruhnya. Hal ini ditujukan agar pemrosesan teks bersifat case-insensitive. Setelah itu, teks berita dilakukan hashing (menggunakan modul HashingTF pyspark) untuk mereduksi menjadi vector-vector yang merepresentasikan teks berita tersebut. Vector tersebut akan dilakukan proses TF (Text Frequency) untuk menghitung jumlah kemunculan suatu teks pada berita tersebut. Setelah dilakukan proses TF, dilakukan proses IDF (Inverse Document Frequency) untuk dibandingkan dengan seluruh text yang terdapat dalam seluruh row.

Program menerima input dari argumen berupa sebuah kata. Kata tersebut lalu akan dicari dokumennya yang paling relevan. Program akan mengoutputkan ID dokumen yang paling relevan dengan kata yang diinputkan tersebut.